ESE Data System Evolution

a.k.a. SEEDS

Kathy Fontaine
GSFC
LP DAAC Meeting
USGS, Reston, VA
September 10, 2003

Topics

- Data System Evolution Goals
- Initial Conditions
- Methodology
- Team Specifics
- What's Next?

Goals

- Given a future paradigm of distributed, heterogeneous data systems and a move toward measurement-focused processing systems, ESE needed
  - A different way of approaching data system management
  - A set of tools and processes to help the ESE make decisions in this environment

Initial Conditions

- Defined in the NewDISS Strategy Document
  - Science questions and priorities must determine the design and function of systems and services.
  - Future requirements will be driven by a high data volume, a broader user base and increasing demand for a variety of data and data products.
  - Technological change will occur rapidly: systems and services must be able to take advantage of these changes.
  - Competition is a key NASA tool for selection of NewDISS components and infrastructure.
  - Principal Investigator (PI) processing and PI data management will be a significant part of future missions and science.
  - Long-term stewardship and archiving must occur.
  - NewDISS evolutionary design must move beyond data and information and towards knowledge-based systems.

Methodology

- NASA invested in a SEEDS study to go from strategic principles to a framework for implementation
  - Seven study team would use community expertise to gather and analyze information.
- All Study Teams' findings are encapsulated in a set of recommendations to ESE, posted July 2003.
  - All Study Teams also recommended working groups as a way to carry out the recommendations.
- Recommendations, responses to comments, and associated technical reports were posted on the web at http://eos.nasa.gov/seeds

Participation

- Study Teams had membership from government, industry, and university
  - Well-represented in the data provider world
  - Representation less from the science user side of the house
- Held three public workshops and many team and stakeholder meetings.

Team Specifics

Levels of Service and Cost Estimation

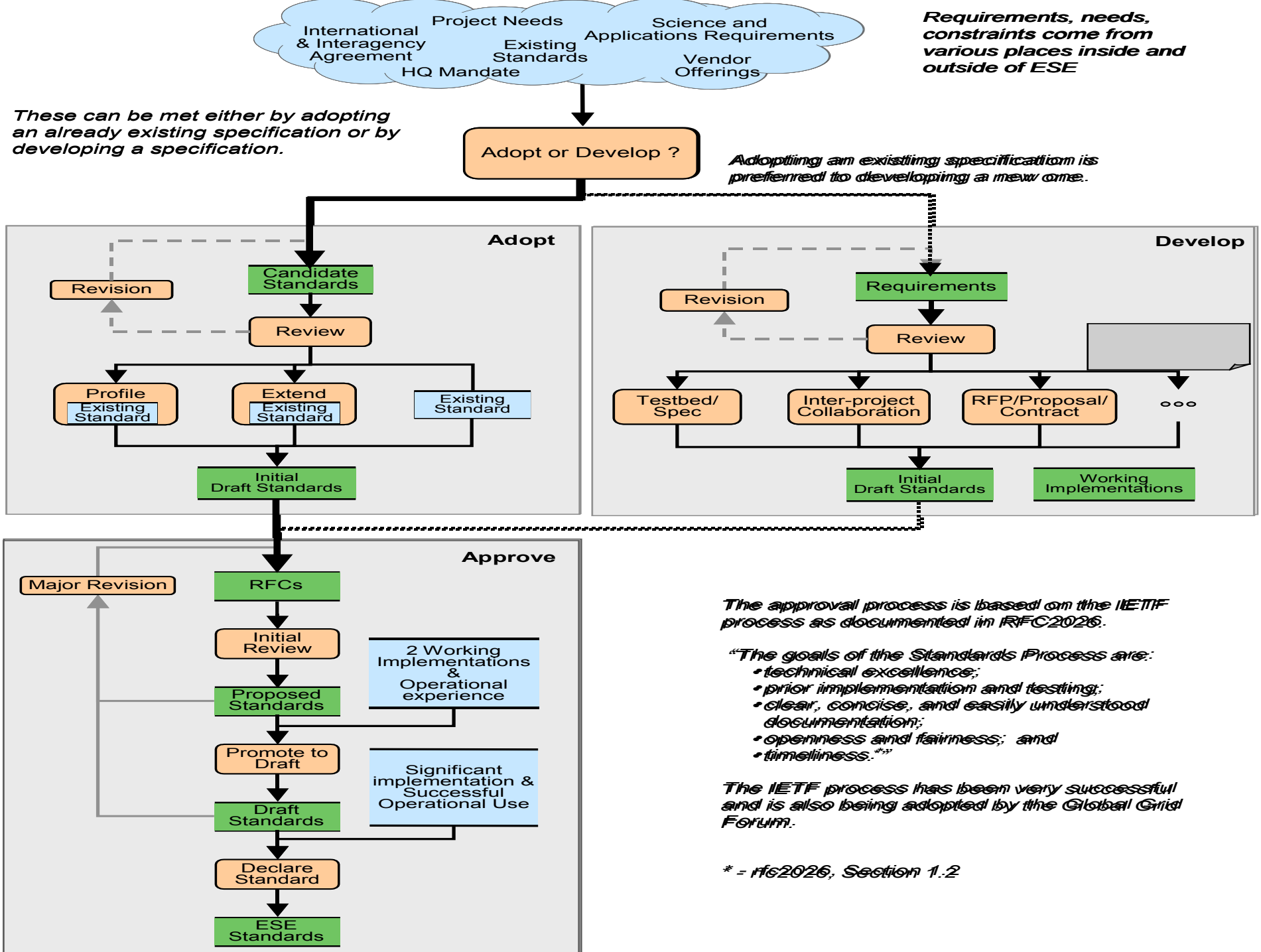
- Purpose
  - Develop baseline levels of service for data service providers; and
  - Develop a cost estimation model that estimates life cycle costs for data service providers.
- Results Include
  - Identified minimum and recommended levels of service for core data sets and services.
  - Created a prototype tool for estimating costs of future data systems based on comparables and levels of effort.

Near-Term Standards and Interfaces

- Purpose
  - To provide specific, concrete recommendations on data format, metadata content, catalog interface, and documentation standards for the near-term missions (LDCM, NPP, OSTM, Ocean Vector Winds, GPM, Solar Irradiance, Carbon Cycle Initiative).
- Results Include
  - Provided a set of mission recommendations and data system standards evolution recommendations.

Long-Term Standards Process

- Purpose
  - Develop a process for ESE to develop, adopt, and evolve standards and interfaces.
- Results Include
  - Realization that there are no 'one-size-fits-all' standards.
  - A process (based on the IETF process) for adopting, approving, or developing ESE standards (process diagram to follow).



The approval process is based on the IETF process as documented in RFC2026.

"The goals of the Standards Process are:

- *technical excellence;*
- *prior implementation and testing;*
- *clear, concise, and easily understood documentation;*
- *openness and fairness; and*
- *timeliness."*

The IETF process has been very successful and is also being adopted by the Global Grid Forum.

** = rfc2026, Section 1.2*

Metrics Planning and Reporting

- Purpose
  - Recommend a governance structure for ESE-funded data systems and services; ensure that metrics planning and reporting processes are commensurate with the needs for accountability; recommend mechanisms to ensure that metrics planning and reporting processes reflect the accomplishments of the data providers and the ESE.
- Results Include
  - Recommendation that metrics be tied to accountability (example of high accountability metrics to follow).

Timeliness Requirement	Description	Sample Metrics
Time-critical, schedule driven operations	All operations schedule-driven; near-real-time critical time constraints; all events scheduled. On-demand production with time constraints. Impact of an operational problem likely to be severe.	Percentage of ingest and production schedules met; Production backlogs; Service interruptions & outages / monthly / trend
Accessibility Requirement	**Description**	**Sample Metrics**
Search and order, data and products access and distribution, services, including user support, are public, open to all users	Services must support large, heterogeneous user community (on the order 10,000 - 100,000), high number of interactions. Problems have wide public exposure.	Profile of user base; Number of accesses; Volume data and products delivered; Volume delivered by request source; User Satisfaction metrics; / monthly / trend
Dependency Requirement	**Description**	**Sample Metrics**
Requires ingest of satellite data streams for product processing; and creates and distributes products required by other DSPs	Ingest of Level 0, or similar satellite data streams; others depend critically on receiving your product(s) in order to perform their functions; performed on an scheduled, operational basis	Percentage of standard products delivered on time to another ESE DSP; Production backlogs; / monthly / trend
Product Quality Requirements	**Description**	**Sample Metrics**
Products generated with peer-reviewed science algorithms; validated, provisional and beta data production supported; robust documentation, quality parameters flagged	Standard products used by users who require science-quality products in their processing and analyses.	Number/List of validated standard products (VSP) generated; Number of standard products cited in literature; Number of distinct users requesting VSP; / monthly / trend
Data Maintenance Requirements	**Description**	**Sample Metrics**
Long-term data stewardship of Level 0 and higher data products received and generated at a DSP	Applicable to long-term data archival facilities where ongoing stewardship is critical to preserving science value of data	Volume of data and products archived by Level; Capacity analysis; Number of accesses of archival data and products > 1 year old; / monthly / trend / media type

Data Life Cycle and Long-Term Archive

- Purpose
  - To develop a set of guidelines to manage ESE data and data needs throughout the data life cycle (cradle to grave).
- Results Include
  - Set of policy guidelines on general, mission, science product generation, active archive, and long term archive responsibilities.

Technology Infusion

- Purpose
  - Define processes to infuse new technologies into the evolving ESE data systems; identify needed capabilities and technologies; determine the roles of ESTO AIST and SEEDS with regard to prototyping needs.
- Results Include
  - Recommendation to adopt and tailor the ESTO AIST process for data systems and services technology needs.
  - Recommendation to define a capability vision for SEEDS.

Reuse and Reuse Architectures

- Purpose
  - Determine if software reuse and reference architectures can reduce the cost and improve the delivery of information services needed by future ESE missions, as well as increase effective and accountable involvement of the community.
- Results Include
  - Defined approaches to reuse and architectures based on community need (mission-critical vs. mission-success).

What's Next?

- Formulation complete
  - NASA will apply the data system evolution framework to existing data and information system elements to determine programmatic evolution direction.
  - Data System Evolution component initiated at GSFC Code 902 (Kathy Fontaine)
    - Four of the six suggested working groups are now being formed
    - REASoN CAN projects will join working groups

Backup

Working Group Activities

- Standards (Richard Ullman)
  - Complete 3 initial RFCs
  - Outreach to missions, projects, science teams
- Metrics Planning and Reporting (H. Ramapriyan)
  - Refine metrics' definitions through MPAR Working Group with REASoN CAN participation
  - Identify common tools/capabilities needed
- …

Working Group Activities

- Technology Infusion (Karen Moe)
  - Identify and prioritize data system technologies to address gaps
  - Identify a plan for acquiring those technologies
- Reuse and Reuse Architectures (Ed Masuoka)
  - Work with mission-critical and mission-success working groups to define how to describe and publish information about reusable assets.

Other Activities

- Form Data Life Cycle Working Group
  - Representation from data providers and users and agency stakeholders.
  - Voluntary effort - no direct funding support.
  - Develop a charter for the DLC Working Group.
    - Identify scope, participants, roles and processes.

Other Activities

- Complete the cost estimation tool and release to the Enterprise at the end of FY04
  - Will include a comparables database of over 20 data activities
  - Will be available on Mac and PC platforms
  - Will be demonstrated early in FY04 to ESE

How to Participate

- To volunteer for any working group
  - Kathy Fontaine
    - Kathy.Fontaine@nasa.gov 301-614-5582
  - Richard Ullman
    - Richard.E.Ullman@nasa.gov 301-614-5228
  - Rama Ramapriyan
    - Rama.Ramapriyan@nasa.gov 301-614-5356
  - Karen Moe
    - Karen.L.Moe@nasa.gov 301-286-2978
  - Ed Masuoka
    - Edward.J.Masuoka@nasa.gov 301-614-5515
  - Ken McDonald
    - Kenneth.R.McDonald@nasa.gov 301-614-5331